

Effects of testing conditions on conceptual survey results

Lin Ding, Neville W. Reay, Albert Lee, and Lei Bao

Department of Physics, The Ohio State University, Columbus, Ohio 43210, USA

(Received 2 March 2008; published 9 June 2008)

Pre-testing and post-testing is a commonly used method in Physics Education Research to assess student learning gains. It is well recognized in the community that timings and incentives in delivering conceptual tests can impact test results. However, it is difficult to control these variables across different studies. As a common practice, a pre-test is often administered either *at* or *near* the beginning of a course, while a post-test can be given either *at* or *near* the end of a course. Also, in conducting such tests there often is no norm as to whether incentives should be offered to students. Because these variations can significantly affect test results, it is important to study and document their impact. We analyzed five years of data that were collected at The Ohio State University from over 2100 students, who took both the pre-test and post-test of the Conceptual Survey of Electricity and Magnetism under various timings and incentives. We observed that the actual time frame for giving a test has a marked effect on the test results and that incentive granting also has a significant influence on test outcomes. These results suggest that one should carefully monitor and document the conditions under which tests are administered.

DOI: [10.1103/PhysRevSTPER.4.010112](https://doi.org/10.1103/PhysRevSTPER.4.010112)

PACS number(s): 01.40.Fk, 01.40.gf

I. INTRODUCTION

Among various educational evaluation techniques,¹ pre-testing and post-testing is the most widely adopted method² in the physics education community. Given the fact that an increasing number of valid and reliable research-based conceptual tests have been developed in the physics domain since the Force Concept Inventory,³ it is now fairly convenient for an instructor to choose a desired test to make a premeasurement and postmeasurement on student conceptual understanding and thus to gauge student learning gains in a particular class. Many physics education researchers have also employed pre and post measurements with these tests to assess the effectiveness of various pedagogical reforms, such as Peer Instruction,^{4,5} cooperative learning,^{6,7} studio physics,^{8–10} Microcomputer Based Laboratory,^{11,12} SCALE-UP (Student Centered Activities in Large Enrollment Undergraduate Program),¹³ modeling instruction,¹⁴ and other interactive engagement pedagogies.^{15–18}

In conducting pre-tests and post-tests, one tacit default is that pre-tests are administered *at* or *near* the beginning of a course, and post-tests are given *at* or *near* the end of the course.¹⁹ It is often expected that student performance will remain approximately the same over a few days at the beginning or end of the class. Additionally, whether or not to grant students incentives for completing these research-based tests often varies across a range of studies. Within a single study, testing conditions including timings and incentives are usually well controlled. However, these conditions may change across different studies, making it difficult to compare results. In existing literature, there is no research on whether and to what extent timings and incentives in delivering conceptual tests may impact test results. A good understanding of this issue is of importance to the Physics Education Research community, particularly as more researchers are starting to collaboratively address similar research questions in different settings.

Over the past five years, the Conceptual Survey of Electricity and Magnetism (CSEM)²⁰ has been administered as

both the pre-test and the post-test in the calculus-based introductory electricity and magnetism (E&M) course at The Ohio State University (OSU). The administration of the CSEM took place at different timings and with various incentives. Continuous use of the CSEM at OSU has so far resulted in a collection of pre-test and post-test matched data from over 2100 students. Based on the analysis of these data, we report findings regarding the effects of test timing and incentives on student performance in the CSEM. The goal of this paper is to provide evidence documenting possible effects of test timings and incentives on student performance, so that researchers can take appropriate controls to address these issues in future studies. In the following, we first present relevant background on the introductory physics course offered and the student populations at OSU (Sec. II); then we report on the analysis of results yielded in different testing conditions. Specifically, we discuss pre-test results at three different timings (before any instruction, after a week of lectures, and after one lecture) and post-test results under four different incentives (no incentives, points for just taking the test, replacing a quiz if scoring high, and part of final examination) (Sec. III). Finally, we discuss implications of the results for future test administration (Sec. IV).

II. BACKGROUND

The calculus-based introductory E&M offered at OSU is the second quarter of the standard introductory physics course for science and engineering majors. Typically, students who attend the E&M classes are mostly freshmen or sophomores, and the majority of them have finished the first quarter of mechanics and met the prerequisite of scoring D or higher.²¹ Materials covered in E&M are standard and include electrostatics, electric circuits, magnetism, and electromagnetic induction. Students meet three times a week for a 48 min lecture delivered by regular faculty in large-lecture halls. Except for summer quarters, typically two different faculty members teach two parallel classes in each quarter.

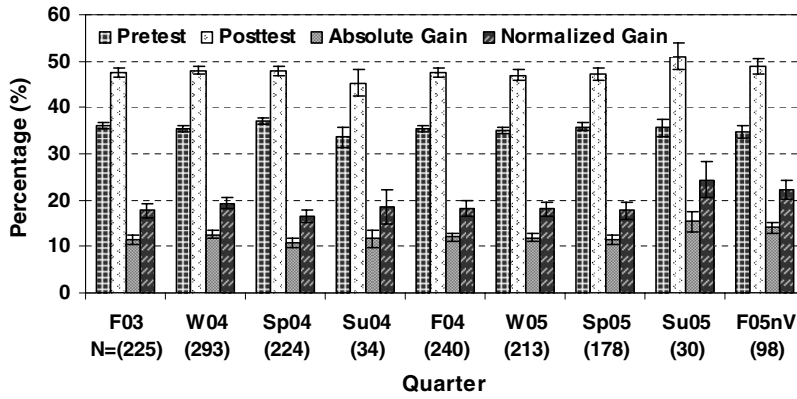


FIG. 1. The pre-test averages, post-test averages, absolute gains, and normalized gains for the individual quarters in the comparison group. (The error bars denote standard errors.)

Before the 2005 Fall quarter, lectures were given traditionally. In the 2005 Fall quarter, one of the authors (N.W.R.) started to adopt in his lecture electronic voting machines (also known as clickers)^{22,23} in combination with various interactive engagement pedagogies. Since then, the clickers have been continuously used in one (and only one) of the two parallel E&M classes of each quarter. For the period from which the data were extracted, students were also required to attend a 48 minute recitation along with a separate 108 min laboratory session each week, both of which are taught by graduate teaching assistants.

III. TEST RESULTS AND ANALYSIS

In the past five years, data from over 2100 students were collected at OSU, and the majority of the data are pre-test and post-test matched. In this paper, we use matched data ($N=2198$) for analysis to trace student performance under various testing conditions (timings and incentives). In the following, we discuss the test results in terms of two major time periods: the quarters from Fall 2003 to Fall 2005 and the subsequent quarters through Spring 2007.

A. Results of the comparison group (Fall 2003–Fall 2005)

From the 2003 Fall quarter through the 2005 Fall quarter, nine different instructors taught the calculus-based introductory E&M course at OSU. Two different textbooks²⁴ and two different homework delivery systems²⁵ also were adopted during this time period. These differences notwithstanding, the course materials covered in class were similar, the laboratories and recitations were essentially the same, and the training of teaching assistants also remained unchanged. In administering the CSEM, the pre-test was always given in the first laboratory of each quarter, which took place during the second school week. Students typically had attended three or more lectures before completing the pre-test. The CSEM post-test was always given in the last laboratory, which was usually conducted in the second-to-last week of a quarter (or, in a few cases, in the last week). No incentives were granted to students for taking either the pre-test or the post-test in any of these quarters. Because the test timings and the lack of incentives were the same for all these quarters, we combine these quarters together and name them the “comparison” group. Results from these quarters set a base-

line for comparison. We have excluded from our analysis one class of the 2005 Fall quarter, in which clicker questions were adopted in lecture. By so doing, we eliminated possible effects from this intervention.

Figure 1 shows the pre-test averages, post-test averages, absolute gains, and normalized gains (see Sec. III D for definitions) of the individual quarters in the comparison group. An ANOVA analysis shows that there is no significant difference across these quarters in the pre-test scores [$F(8, 1526)=0.84$, $p=0.5678$], post-test scores [$F(8, 1526)=0.43$, $p=0.9052$], or gains [absolute gains: $F(8, 1526)=0.91$, $p=0.5097$; normalized gains: $F(8, 1526)=0.95$, $p=0.4759$]; thus, confirming the validity of combining these quarters as a comparison group.

B. Pre-test results of the subsequent quarters (Winter 2006–Spring 2007) and comparisons with the comparison group

In subsequent quarters from Winter 2006 through Spring 2007, the CSEM pre-test was given under two different timings: on the first day of class (either in lecture or in recitation) or after one lecture. Quarters of 2006 Winter and 2007 Spring belong to the former case, and the 2006 Spring quarter to the latter. Similarly to the comparison group, no incentive was offered in any of these quarters. We combine the 2006 Winter and 2007 Spring quarters together for analysis and label them the “no-instruction” group. (Here we did not exclude the data of the clicker classes, as no intervention had been introduced prior to the pre-test.) In the following, we discuss the pre-test results of the no-instruction group and the 2006 Spring quarter and compare these results with those

TABLE I. Pre-test conditions and results.

Quarter (No. of students)	Pre-test average \pm Std. error	Timing and incentive
Comparison group ($N=1535$)	11.4 ± 0.1	After one week/no incentives
No-instruction group ($N=563$)	9.0 ± 0.1	First day/no incentives
2006 Spring ($N_{\text{nonclicker}}=100$; $N_{\text{clicker}}=54$)	9.0 ± 0.3 (nonclicker) 10.8 ± 0.5 (clicker)	After one lecture/no incentives

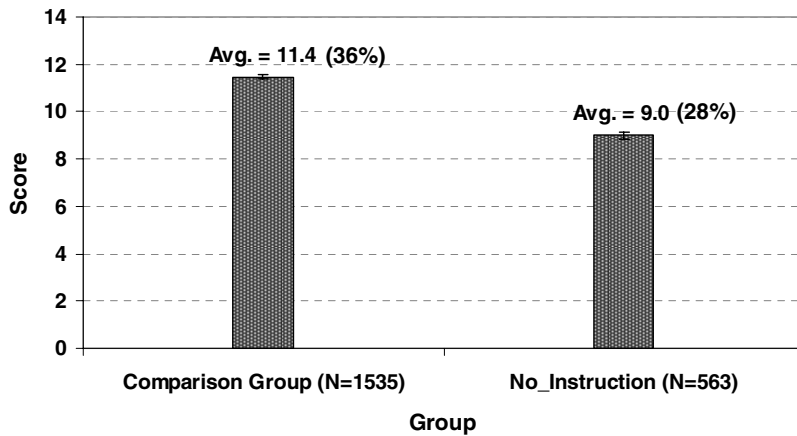


FIG. 2. Pre-test total scores of the comparison group and the no-instruction group. (The percentages indicate average pre-test score percentages; the error bars denote standard errors.)

of the comparison group. We show that pre-test results can be significantly affected by a week of lectures or sometimes even a single lecture. Table I shows the pre-test conditions and results.

Figure 2 displays the pre-test total scores of the comparison group and the no-instruction group. The comparison group outperformed the no-instruction group by 8%, which is equivalent to two and half questions. A t -test suggests that the difference in pre-test score between the two groups is both significant ($t=13.63$, $p<0.001$) and sizable (effect size=0.7).²⁶ This result indicates that if the CSEM pre-test is conducted a few days and lectures into a course, student overall performance is noticeably better than that if the pre-test is conducted before any instruction takes place.

Our analysis further shows that the better performance of the comparison group in the CSEM pre-test is mainly from their higher scores on the electricity questions (Q1–Q20). However, the comparison group did not outperform the no-instruction group on all the electricity questions. From Fig. 3 where the individual item scores are plotted, we find the pre-test difference between the two groups lies mostly in the first nine questions. Note that these nine questions mainly deal with “electric charge and force,”²⁷ which are exactly the topics discussed in the first several lectures of a quarter. For these questions, the average difference (Δ_1) between the two groups is 20%, equivalent to two questions, which accounts for a large percentage of the difference detected in the total

score. On the other hand, for the remaining electricity questions that address “electric field and force” or “electric potential and energy”²⁷ (topics not covered in the first week), the average difference (Δ_2) is only 6%, equivalent to half a question. (One clarification worth making is that the curves in Fig. 3 do not intend to imply continuous data but rather to provide a better visual effect on the trend of item scores.)

The above results suggest that a week of lectures can have a significant effect on pre-test results. As a matter of fact, we find that sometimes even one lecture can markedly impact pre-test results depending on what is covered in that lecture. In the 2006 Spring quarter, where the CSEM pre-test was administered in a recitation after the first lecture without incentives, one instructor (in the clicker class) unknowingly discussed several CSEM questions in his first lecture. As a result, the average pre-test score for that class turned out to be 10.8, similar to that of the comparison group ($t=1.16$, $p=0.2442$). Conversely, the other instructor (in the nonclicker class) spent nearly half of the class time addressing logistic issues and covered less material in the first lecture. Consequently, that class only scored average 9.0, noticeably lower than that of the comparison group ($t=6.53$, $p<0.001$). Clearly, depending on what is covered in one lecture, the impact of that lecture on pre-test results sometimes cannot be ignored.

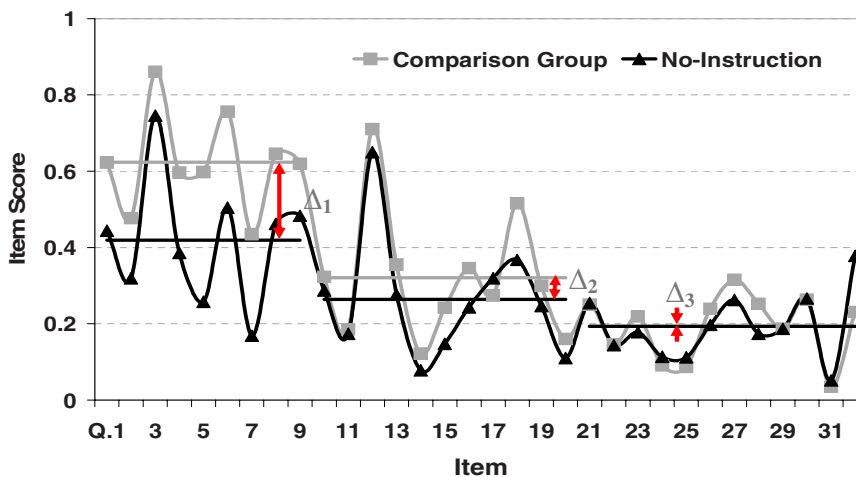


FIG. 3. (Color) Pre-test item scores of the comparison group and the no-instruction group. The curves show the trend of item scores; the straight lines connect the average item scores in each range.

TABLE II. Post-test results and conditions.

Quarter (No. of students)	Post-test average \pm Std. error	Timing and incentive
Comparison group ($N=1535$)	15.2 ± 0.1	Last laboratory/no incentives
2006 Winter ($N=175$)	15.0 ± 0.5	Last laboratory/small amount of points for just taking the test
2006 Spring ($N=100$)	17.6 ± 0.5	Last recitation/score $>90\%$ to replace a lowest quiz score
2007 Spring ($N=121$)	19.9 ± 0.4	Final exam

C. Post-test results of the subsequent quarters and comparisons with the comparison group

Table II lists the post-test conditions (timings and incentives), with results for the comparison group and the individual quarters from 2006 to 2007. Since the post-test conditions were all different from 2006 to 2007, we discuss each quarter separately. In the analysis of the post-test results, we have retained the data only from the nonclicker classes to eliminate possible effects from the “clicker” intervention.

In the 2006 Winter quarter, the post-test was administered with the same timing as in the comparison group (in the last laboratory) but with an incentive. Students who completed the post-test would get a small amount of points regardless of how they performed on the test. Consequently, a large fraction of the class took the post-test and resulted in nearly 90% pre-test and post-test matched data of the entire class, higher than that of the comparison group (average of 72%). However, the post-test average was only 15, slightly but not significantly lower than that of the comparison group [$t=0.71$, $p=0.4771$] (see Fig. 4). Possibly, the kind of incentive offered in this quarter had drawn a larger fraction of the class to take the post-test, including those lesser-achieving students, which in turn yielded a slightly lower average than that of the comparison group.

In the 2006 Spring quarter, the post-test was administered during the last recitation, which took place several days after the last laboratory. Another type of incentive was offered; students were told that if they scored 90% or higher, they could replace the CSEM score for a lowest quiz score. The participation rate in that quarter dropped significantly; the

percentage of students taking both the pre-test and post-test was only 65%. However, the post-test average was noticeable higher compared with the comparison group ($t=4.58$, $p<0.0001$; effect size=0.5). Using a scale that goes from 4 for grade A (excellent) to 0 for grade E (fail), we found that students who took both the pre-test and the post-test obtained an average of 2.63 in the course final grade, whereas those who missed at least one of these tests had an average of 1.86. The post-test incentive offered in the 2006 Spring quarter may have attracted only more motivated and achieving students to take the post-test, increasing the post-test score.

In the 2007 Spring quarter, the post-test CSEM was incorporated into the final exam. Students were able to review course materials before taking the test and were motivated to answer the questions correctly. Our results show that the post-test average is 19.9, the highest among all the quarters even with more than 90% of all students participating. Compared to the comparison group, the increase in the post-test score is both significant and large ($t=9.73$, $p<0.0001$; effect size=0.9).

These analyses illustrate possible effects of testing timings and incentives on test outcomes. Particularly, different incentives seem to attract different fractions of a class to complete the test, which may cause a noticeable fluctuation in the results.

D. Gains and normalized gains of the subsequent quarters and comparisons with the comparison group

In gauging the change of student performance after course instruction, absolute gain and normalized gain²⁸ are perhaps

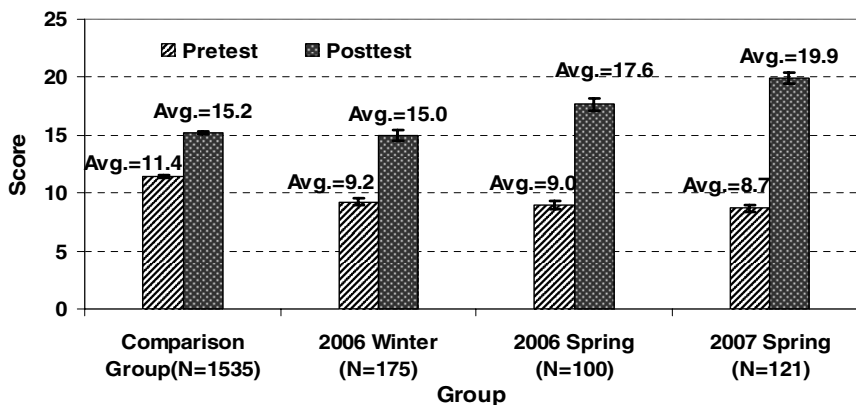


FIG. 4. Pretotal and Post-total scores of the comparison group and the subsequent quarters. Note that the pre-test scores of the subsequent quarters are similar, but the post-test scores are rather different under different test timings and incentives. (The error bars denote standard errors.)

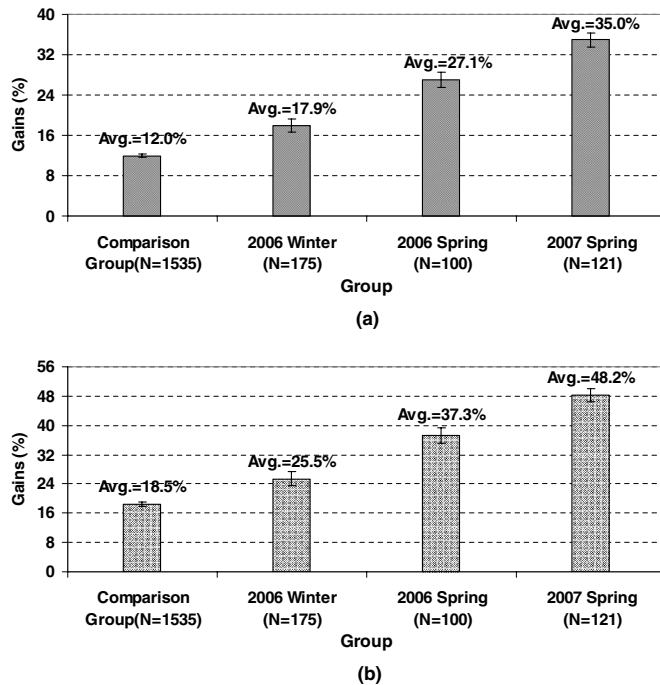


FIG. 5. (a) Absolute gains of the comparison group and the subsequent quarters. Note that the pre-test scores for quarters from 2006 to 2007 are rather similar (see Fig. 4). (b) Normalized gains of the comparison group and the subsequent quarters. Note that the error bars denote standard errors.

the two most commonly used measures. They are expressed respectively as follows:

$$\text{Absolute gain} = (\text{postscore}) \% - (\text{prescore}) \%,$$

$$\text{Normalized gain} = \frac{(\text{postscore}) \% - (\text{prescore}) \%}{100 \% - (\text{prescore}) \%}.$$

In the following, we use both measures to demonstrate the effects of test timing and incentive on the CSEM test outcomes.

Figure 5(a) displays the absolute gains of the comparison group and the subsequent quarters. By adjusting pre-test and

post-test timings and/or incentives, we have observed an absolute gain from as low as 12% (equivalent to 4 questions) up to 35% (equivalent to 11 questions). Normalized gains are given in Fig. 5(b). It is evident that normalized gains have increased from 18.5% in the comparison group up to 48.2% in the 2007 Spring quarter.

IV. DISCUSSION AND IMPLICATIONS

Although one can administer a pre-test either at or near the beginning of a course, our results suggest that the CSEM pre-test scores are sensitive to moving the test even a few days and lectures. Similarly, when to administer a post-test also has a significant effect on test results. Besides, incentives also have a potential impact on student performance; different incentives may attract different fractions of students to take the post-test, impacting test outcomes.

It follows that absolute or normalized gains may also vary greatly. In our analysis of the data that were collected in the past 12 quarters over five years, normalized gains for a traditionally taught course varied from 18.5% to 48.2%. Note that when pre-test and post-test conditions were maintained consistently in the comparison group, years of data showed a fairly stable normalized gain $18.5\% \pm 0.6\%$ (std. error). Table III summarizes the results.

Although it is widely accepted that different timings and incentives may impact test results, the extent of this impact is still largely unclear within the physics education community. To this end, we present the above results in the hope of alerting instructors and researchers to the potentially large effects of test timings and incentives on student performance and test outcomes. We encourage interested readers to further investigate how the analysis and results based on the CESM data collected at OSU extrapolate to other institutions, student populations, and conceptual tests.

ACKNOWLEDGMENTS

The authors wish to thank the National Science Foundation under Grant No. DUE 0618128 and also The Ohio State University for providing funding in support of this research.

TABLE III. A summary of the test conditions and test results.

Group	Pre-test conditions	Post-test conditions	Preaverage	Postaverage	Normal Gain (%)
Comparison Group (N=1535)	Laboratory in second week; no incentive	Last laboratory; no incentive	11.4	15.2	18.5
2006 Winter (N=175)	First day; no incentive	Last laboratory; a few points for just taking the test	9.2	15.0	25.5
2006 Spring (N=100)	After one lecture; no incentive	Last recitation; score >90% replaces a lowest quiz score	9.0	17.6	37.3
2007 Spring (N=121)	First day; no incentive	Final exam	8.7	19.9	48.2

- ¹J. R. Fraenkel and N. E. Wallen, *How to Design and Evaluate Research in Education*, 3rd ed. (McGraw-Hill, New York, 1996).
- ²M. D. Sundberg, Assessing student learning, *Cell Biol. Educ.* **1**, 11 (2002); R. R. Hake, Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.* **66**, 64 (1998); L. Bao and E. F. Redish, Model analysis: Representing and assessing the dynamics of student learning, *Phys. Rev. ST Phys. Educ. Res.* **2**, 010103 (2006); L. Bao, Theoretical comparisons of average normalized gain calculations, *Am. J. Phys.* **74**, 917 (2006); J. D. Marx and K. Cummings, Normalized change, *ibid.* **75**, 87 (2007).
- ³D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, *Phys. Teach.* **30**, 141 (1992); R. Thornton and D. Sokoloff, Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula, *Am. J. Phys.* **66**, 338 (1998); R. Beichner, Testing student interpretation of kinematics graphs, *ibid.* **62**, 750 (1994); L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, *Phys. Rev. ST Phys. Educ. Res.* **2**, 010105 (2006); P. V. Engelhardt and R. J. Beichner, Students' understanding of direct current resistive electrical circuits, *Am. J. Phys.* **72**, 98 (2004), and many others.
- ⁴A. P. Fagen, C. H. Crouch, and E. Mazur, Peer Instruction: Results from a range of classrooms, *Phys. Teach.* **40**, 206 (2002).
- ⁵C. H. Crouch and E. Mazur, Peer Instruction: Ten years of experience and results, *Am. J. Phys.* **69**, 970 (2001).
- ⁶D. Johnson, R. Johnson, and K. Smith, Cooperative Learning: Increasing college faculty instructional productivity, ASHE-ERIC Higher Education Report No. 4, 1991.
- ⁷M. Samiullah, Effect of in-class student-student interaction on the learning of physics in a college physics course, *Am. J. Phys.* **63**, 944 (1995).
- ⁸J. Wilson, The CUPLE physics studio, *Phys. Teach.* **32**, 518 (1994).
- ⁹K. Cummings, J. Marx, R. Thornton, and D. Kuhl, Evaluating innovation in studio physics, *Am. J. Phys.* **67**, S38 (1999).
- ¹⁰C. Hoellwarth, M. J. Moelter, and R. D. Knight, A direct comparison of conceptual learning and problem solving ability in traditional and studio style classrooms, *Am. J. Phys.* **73**, 459 (2005).
- ¹¹R. K. Thornton and D. R. Sokoloff, Learning motion concepts using real-time microcomputer-based laboratory tools, *Am. J. Phys.* **58**, 858 (1990).
- ¹²E. F. Redish, J. M. Saul, and R. N. Steinberg, On the effectiveness of active-engagement microcomputer-based laboratories, *Am. J. Phys.* **65**, 45 (1997).
- ¹³R. Beichner, J. M. Saul, D. S. Abbott, J. Morse, Duane Deardorff, Rhett J. Allain, S. W. Bonham, Melissa Dancy, and J. Risley, Student-Centered Activities for Large Enrollment Undergraduate Programs (SCALE-UP) project, in *Research-Based Reform of University Physics*, edited by E. F. Redish and P. J. Cooney (American Association of Physics Teachers, College Park, MD, 2006).
- ¹⁴D. Hestenes, Findings of the modeling workshop project (1994–2000) (retrieved from <http://modeling.asu.edu/R&M/ModelingWorkshopFindings.pdf>).
- ¹⁵R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.* **66**, 64 (1998).
- ¹⁶R. Beichner, L. Bernold, E. Burniston, P. Dail, R. Felder, J. Gastineau, M. Gjertsen, and J. Risley, Case study of the physics component of an integrated curriculum, *Am. J. Phys.* **67**, S16 (1999).
- ¹⁷R. N. Steinberg and K. Donnelly, PER-based reform at a multicultural institution, *Phys. Teach.* **40**, 108 (2002).
- ¹⁸D. E. Meltzer and K. Maivannan, Transforming the lecture-hall environment: The fully interactive physics lecture, *Am. J. Phys.* **70**, 639 (2002).
- ¹⁹If a chosen test covers only a few specific topics, then the pre-test and post-test timings are relaxed, respectively, to be before and after relevant materials being discussed in class.
- ²⁰D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, *Am. J. Phys.* **69**, S12 (2001).
- ²¹Edward Adelson (private communication).
- ²²N. W. Reay, L. Bao, P. Li, R. Warnakulasooriya, and G. Baugh, Toward the effective use of voting machines in physics lectures, *Am. J. Phys.* **73**, 554 (2005).
- ²³N. W. Reay, P. Li, and L. Bao, Testing a new voting machine question methodology, *Am. J. Phys.* **76**, 171 (2008).
- ²⁴R. D. Knight, *Physics for Scientists and Engineers*, 1st ed. (Addison-Wesley, San Francisco, 2004); and D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics*, 7th ed. (Wiley, New York, 2004); Edward Adelson (private communication).
- ²⁵Thomas Bolland, private communication; For information on WebAssign, see J. D. Risley, WebAssign: Assessing student performance any time any where, Uniserve Science News, 13 (available at <http://science.uniserve.edu.au/newsletter/vol13/risley.html>) and Mastering Physics; For information on Mastering Physics, see D. Pritchard, Mastering Physics, Pearson Education, 2004 (available at <http://www.masteringphysics.com/>).
- ²⁶J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Lawrence Earlbaum, Hillsdale, NJ, 1988).
- ²⁷M. Planinic, Assessment of difficulties of some conceptual areas from electricity and magnetism using the Conceptual Survey of Electricity and Magnetism, *Am. J. Phys.* **74**, 1143 (2006).
- ²⁸F. W. Gery, Does mathematics matter?, in *Research Papers in Economic Education*, edited by Arthur Welsh (Joint Council on Economic Education, New York, 1972).